Yihao (Q) Quan

E-mail: qyhhere@gmail.com | Tel: +1 848-247-9940

Personal Page | Github | Google scholar | Linkedin

Education

- **Rutgers University - New Brunswick** *Sept 2025 - Present*
  Ph.D. Student in Computer Science

- **Beijing Jiaotong University** *Sept 2021 - July 2025*
  B.Sc. in Management Information Systems

Research Interests

My primary research lies in **MLLMs/LLMs**. Specifically:

- **Vision + Language:** Bridging the gap between language and vision representations, with the goal of building more interpretable and controllable multimodal systems.
- **Hallucination, Factuality, and Safety:** Leveraging interpretability insights to improve downstream capabilities such as factuality, reasoning, hallucination mitigation, and model editing, while also designing safer models.
- **Agentic AI:** Developing agentic AI systems with reliable memory, adaptive planning, and joint learning between agents and world models.

Selected Publications

*** Equal Contribution**

- **[COLM 2026] Reinforcing Consistency in Video MLLMs with Structured Rewards**
  MLLMs Video Understanding
  Yihao Quan, Zeru Shi, Jinman Zhao, Ruixiang Tang.
  arXiv
- **[EMNLP 2025, *Oral, 325/8174 ≈ 3.9%*] Shallow Focus, Deep Fixes: Enhancing Shallow Layers Vision Attention Sinks to Alleviate Hallucination in LVLMs**
  MLLMs Hallucination
  Yihao Quan*, Xiaofeng Zhang*, Chaochen Gu, Chen Shen, Xiaosong Yuan, Hao Cheng, Kaijie Wu, Shaotian Yan, Jieping Ye.
  Code arXiv
- **[NAACL 2025, *Oral, 226/3332 ≈ 6.7%*] From Redundancy to Relevance: Information Flow in LVLMs Across Reasoning Tasks**
  MLLMs Visual Reasoning Interpretability Efficiency
  Xiaofeng Zhang*, **Yihao Quan***, Chen Shen, Chaochen Gu, Xiaosong Yuan, Shaotian Yan, Liang Xie, Wenxiao Wang, Hao Tang, Jieping Ye.
  Code arXiv
- **[ACL 2026 Main, *Oral Recommended*] ART: Attention Replacement Technique to Improve Factuality in LLMs**
  LLMs Factuality
  Ziqin Luo*, **Yihao Quan***, Xiaofeng Zhang, Xiaosong Yuan, Chen Shen.
  Code arXiv
- **[ECCV 2026] Counting Circuits: Mechanistic Interpretability of Visual Reasoning in Large Vision-Language Models**
  MLLMs Visual Reasoning Interpretability
  Liwei Che*, Zhiyu Xue*, **Yihao Quan***, Benlin Liu, Zeru Shi, Ruixiang Tang, Ranjay Krishna, Vladimir Pavlovic.
  arXiv
- **[NeurIPS 2026] AgentLens: Interpretable Safety Steering via Mechanistic Subspaces for Multi-Turn Coding Agent**
  Agentic Safety Interpretability
  Weidi Luo, Qiming Zhang, **Yihao Quan**, Mingyu Jin, Jie Cai, Chaowei Xiao, Jingcheng Niu, Zhen Xiang.
  Code arXiv

- **[NeurIPS @ ResponsibleFM] Read the Scene, Not the Script: Outcome-Aware Safety for LLMs**
  LLMs Safety Interpretability
  Rui Wu, **Yihao Quan**, Zeru Shi, Zhenting Wang, Yanshu Li, Ruixiang Tang.
  Code arXiv
- **[AAAI 2025] Enhancing Multimodal Large Language Models Complex Reason via Similarity Computation**
  MLLMs Visual Reasoning Efficiency
  Xiaofeng Zhang*, Fanshuo Zeng*, **Yihao Quan**, Zheng Hui, Jiawei Yao.
  Code arXiv

Selected Preprints

* Equal Contribution

- **[In Submission] Preserve Content, Disrupt Structure: Patch-Shuffle Counterfactual Distillation for Visual Reasoning**
  MLLMs On Policy Distillation
  Yihao Quan, Jiang Liu, Ximeng Sun, Jialian Wu, Ruixiang Tang, Emad Barsoum, Zicheng Liu.
  arXiv
- **[In Submission] Why MLLMs Struggle to Count: Overcoming Individuation and Aggregation Bottlenecks with ConvStack**
  MLLMs Visual Reasoning Interpretability
  Liwei Che*, **Yihao Quan***, Sen Fang, Hongyi Wang, Ranjay Krishna, Ruixiang Tang, Vladimir Pavlovic.
  arXiv
- **[In Submission] Improving Visual Reasoning with Iterative Evidence Refinement**
  MLLMs Visual Reasoning Thinking with Image
  Zeru Shi*, Kai Mei*, **Yihao Quan**, Dimitris N. Metaxas, Ruixiang Tang.
  arXiv
- **[In Submission] MemEye: A Visual-Centric Evaluation Framework for Multimodal Agent Memory**
  Agentic Safety Interpretability
  Minghao Guo*, Qingyue Jiao*, Zeru Shi*, **Yihao Quan**, Boxuan Zhang, Danrui Li, Liwei Che, Wujiang Xu, Shilong Liu, Zirui Liu, Mubbasir Kapadia, Vladimir Pavlovic, Jiang Liu, Mengdi Wang, Yiyu Shi, Dimitris N Metaxas, Ruixiang Tang.
  Code arXiv
- **[In Submission] The Anatomy of a Fake Citation: Field-Specific Hallucination Neurons in Large Language Models**
  LLMs Hallucination Interpretability
  Yuefei Chen, **Yihao Quan**, Xiaodong Lin, Ruixiang Tang.
  arXiv

Selected Projects

- **Awesome-LMMs-Mechanistic-Interpretability (*Projects Lead*)** *200+ stars*
  A curated collection of resources focused on the Mechanistic Interpretability (MI) of Large Multimodal Models.
- **Awesome-World-Models (*Projects Contributor*)** *3000+ stars*
  A curated collection of resources focused on World Models (WMs).

Industry Experience

* Publications Listed Include Only (Co-)First Author Works

- **AMD GenAI Team** *Jun 2026 – Present*
  Research Topic: Multimodal On Policy Distillation
  Advisor: Jiang Liu
  Preprint: In Submission
- **Alibaba** *Jun 2024 – Sep 2024*
  Research Topic: Vision Hallucination Mitigation
  Mentor: Chen Shen & Xiaofeng Zhang
  Publication: EMNLP'25 (Oral)

Research Experience

*** Publications Listed Include Only (Co-)First Author Works**

- **Rutgers University** *Sep 2025 – Present*
  Research Topic: Video Understanding Enhancement, Mechanistic Interpretability
  Advisor: Ruixiang Tang
  Publication: In Submission, COLM'26, ECCV'26

- **KAUST (King Abdullah University of Science and Technology)** *Sep 2024 – Nov 2024*
  Research Topic: Inefficient Attention Allocation in LVLMs
  Mentor: Yang Shu

- **Shanghai Jiao Tong University** *Feb 2024 – Jun 2024*
  Research Topic: Information Flow in LVLMs
  Mentor: Xiaofeng Zhang
  Publication: NAACL'25 (Oral), ACL'26 (Main)

Professional Service

Reviewer: ICLR'27, COLM'26, ICLR'25, AAAI'26, AAAI'27, ARR Oct, ARR July, ARR May

Selected Awards And Honors

- **COLM'26 Travel Award** *2026*
- **SIAM'26 Travel Award** *2026*